

# Discovering Babel: enhanced language resource discovery

**Jens Stegmann**

*Institut für Deutsche Sprache*

&

*Institut für Maschinelle*

*Sprachverarbeitung, Universität  
Stuttgart*

[jens.stegmann@gmail.com](mailto:jens.stegmann@gmail.com)

**Martin Wynne**

Head of the Oxford Text Archive

*Oxford University Computing Services,*

*Oxford e-Research Centre &*

*Faculty of Linguistics, Philology and Phonetics*

University of Oxford

[martin.wynne@oucs.ox.ac.uk](mailto:martin.wynne@oucs.ox.ac.uk)



**Oxford University Computing Services**

[www.oucs.ox.ac.uk](http://www.oucs.ox.ac.uk)

# The problems facing users of language resources

---

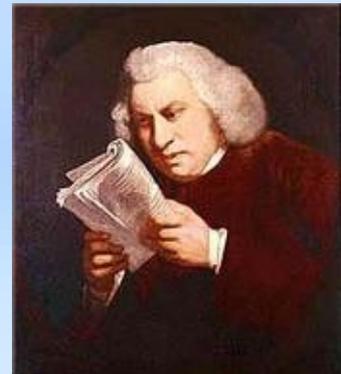
- Many archives known only to certain communities
- Archives are mostly unconnected, and data difficult to find
- Every archive has its own standards for storage and access
  - usually only simple retrieval of files (text, audio or video documents)
- Not sufficient incentives to share resources
- Resources are in different formats, follow different standards, are described in differing ways
- Tools are hard to use for non-specialist
- Tools and data are not available for online processing
- Many researchers are not aware of the potential benefits of using language and speech technology tools



# Challenges for creators of resources

---

- **Standards**
- **Interoperability**
- **Sustainability**
- **Connecting and linking:** how can I facilitate research using my data along with other ones?
- **Creating a finished product:** when should I stop developing and make it available?
- **Impact:** how can I create the maximum impact with my data, in my academic field and beyond?
- How do I ensure that my resource is made available in relevant ways ***as a service***



 Oxford University Computing Services

Thursday 23. Jun 2011

Oxford Text Archive: [Home](#) | [About](#) | [News](#) | [Catalogue](#) | [Contact](#) | [Help and FAQ](#) | [Search OTA](#)

Please note that some of our resources do not appear in this table. Reasons for this might include that they were deposited with us for preservation only, and not dissemination, or that our legal right to disseminate them has been questioned.

ID	Availability	Title	Language	Author
<a href="#">2541</a>	free	VU Amsterdam Metaphor Corpus	English	Gerard J Steen; Aletta G Dorst; J Berenike Herrmann; Anna A Kaal; Tina Krennmayr
<a href="#">2540</a>	free	Speech, Thought and Writing Presentation Corpus (STWP)	English	Culpeper, Jonathon; Semino, Elena; Short, Mick; Wynne, Martin
<a href="#">2539</a>	restricted	British Academic Written English Corpus	English	Nesi, Hilary; Gardner, Sheena; Thompson, Paul; Wickens, Paul
<a href="#">2537</a>	free	GerManC. A Historical Corpus of German Newspapers 1650-1800	German	Durrell, Martin; Ensslin, Astrid; Bennett, Paul (ed.)
<a href="#">2531</a>	free	The Lancaster Newsbooks Corpus	English	Thomason, George, d. 1666
<a href="#">2530</a>	restricted	Language convergence and grammatical borrowing database	English	
<a href="#">2529</a>	restricted	The Workdiaries of Robert Boyle	English	Hunter, Michael; Centre for Editing Lives and Letters
<a href="#">2528</a>	free	Demetrios Database of Septuagint Greek	English	
<a href="#">2527</a>	restricted	Chambers-Le Baron Corpus of Research Articles in French	French	Chambers, Angela; Le Baron, Florence
<a href="#">2525</a>	restricted	British Academic Spoken English corpus	English	Nesi, Hilary; Thompson, Paul
<a href="#">2524</a>	free	Discourse on the Origin and the Foundations of Inequality Among Men	English	
<a href="#">2523</a>	free	The Birth of Tragedy	English	
<a href="#">2522</a>	free	On the Use and Abuse of History for Life	English	
<a href="#">2521</a>	free	Universal Natural History and Theory of Heaven	English	
<a href="#">2520</a>	free	Cognition, Biology and Idealist Philosophy	English	Randrup, Axel
<a href="#">2518</a>	restricted	The Electronic Text Corpus of Sumerian Literature. Revised edition.	English	Cunningham, Graham; Ebeling, Jarle; Black, Jeremy (deceased); Flückiger-Hawker, Esther; Robson, Eleanor; Taylor, Jon; Zólyomi, Gábor (ed.)
<a href="#">2517</a>	restricted	Angloromani (sample)	English	Matras, Yaron
<a href="#">2516</a>	free	Discourse context and the processing of contrastive focus in silent reading (SPSS data files)	English	Paterson, Kevin
<a href="#">2514</a>	free	Correspondences: Jewish Mysticism, Indian Philosophies	English	Randrup, Axel; Bagchi, Tista

# Discovering Babel

---

## **Discovering Babel: Enhanced Language Resource Discovery**

*“The digital literary and linguistic resources in the Oxford Text Archive and in the British National Corpus have been available to researchers throughout the world for several decades. Technical enhancements to the resource discovery infrastructure will allow wider dissemination of open metadata, will facilitate interaction with research infrastructures, and the knowledge and expertise achieved will be shared with the community.”*

**<http://www.jisc.ac.uk/whatwedo/programmes/inf11/infrastructureforresourcesdiscovery/discoveringbabel.aspx>**

# Discovering Babel

---

- c. 1400 metadata records;
- c. 1400 electronic literary and linguistic datasets:
  - Electronic texts
  - Text corpora
  - Lexicons
  - Databases of linguistic information
  - Audio data
- British National Corpus

# Sharing metadata

---

## Resource discovery metadata:

- Text Encoding Initiative (TEI) XML headers
- Dublin Core;
- Open Language Archives Community (OLAC) format (extended DC);
- CLARIN Metadata Initiative (CMDI)
- RDF linked data;
- OAI-PMH target
- Harvested by OLAC, CLARIN, etc.



# OLAC Language Resource Catalog

Search for language resources  go

- Navigating the Catalog**
  - Catalog Home
  - Search Strategies
  - Advanced Search
  - New: Records recently added or modified
- Quick Links**
  - Browse by Language
  - Browse by Country
  - Browse by Linguistic Field
  - Browse by Linguistic Type
  - Browse by Language Family
- Contacts**
  - Email Us
- More information**
  - OLAC Homepage
  - OLAC FAQ
  - Participating Archives

Powered by the DLA

**Results:** [« First](#) • [Previous](#) • [Next](#) • [Last »](#)

Showing hits **1 - 50** out of **1264**

**Sectio canonis**  
None Specified. 1911-01-01. Oxford Text Archive.

**The second part of the bloody conquests of mighty Tamburlaine : with his impassionate fury for the death of his lady and love, fair Zenocrate, his form of exhortation and discipline to his three sons, and the manner of his own death**  
Marlowe, Christopher. 1979-09-08. Oxford Text Archive.

**Quintessence of Ibsenism**  
Shaw, Bernard. 1993-05-05. Oxford Text Archive.

**Fanny Hill : memoirs of a woman of pleasure**  
Cleland, John. 1993-05-25. Oxford Text Archive.

**Vita Davidus**  
Rhygyfarch. 1967-01-01. Oxford Text Archive.

**The arte of rhetorique : for the use of all suche as are studious of eloquence, sette forth in English**  
Wilson, Thomas. 1996-02-23. Oxford Text Archive.

**A woman killed with kindness**  
Heywood, Thomas, d. 1991-07-10. Oxford Text Archive.

**King Lear : 1608**  
Shakespeare, William. 1991-04-12. Oxford Text Archive.

**De cive : containing the elements of civill politie in the agreement which it hath both with naturall and divine lawes in which is demonstrated, both what the origine of justice is, and wherein the essence of Christian religion doth consist together with the nature, limits and qualifications both of regiment and subjection**  
Hobbes, Thomas. 1996-05-17. Oxford Text Archive.

**Morgante**  
Pulci, Luigi. 1989-12-08. Oxford Text Archive.

**The saga of Grettir the Strong**  
None Specified. 1996-01-18. Oxford Text Archive.

**The Ion of Euripides**

**Currently Used Filters**

- Archive: Oxford Text Archive

**Sort Results By:**

**Possible Sorts:**

all	
Title	[a-z][z-a]
Id	[a-z][z-a]
Date	[a-z][z-a]

**Narrow Results By:**

**Format** browse

• text/plain	1085
• application/sgml	154
• application/gtar	7
• application/gzip	6

[view more...](#)

**Contributor** browse

• None Specified	114
• Not Available	102
• Triggs, Jeffery	63
• Shakespeare, William	55

[view more...](#)

**Title** browse

• 1984	5
• Dubliners	4
• Volpone 1607	4
• A portrait of the artist as a young man	3

[view more...](#)

**Other date** browse

• 1911-01-01	183
• 1976-01-01	75
• 1993-06-08	43
• 1993-06-10	43

[view more...](#)

**Other format** browse

# Key challenges

---

- Establishing sensible and standards-conformant architecture for resource file locations;
- Conformance to semantics of various target metadata schemas;
- Expressing quality assurance metadata for legacy data;
- Expressing information for web services processing;
- Mapping licence restrictions to CLARIN 'laundry symbols';
- Establishing procedures for ensuring persistence and high availability of services.

# Discovering Babel

---

## How will the data be made available?

- All data continues to be available for download via the OTA catalogue;
- Establish permanent, highly available online locations for all data files;
- Some data open access, some behind Shibboleth authorization, some requiring manual authorization;
- Register handles for persistent locations with EPIC;
- Key datasets to be hosted on National Grid Service (NGS);
- Combination with OxGarage text conversion service allows download of various formats, e.g. ePub.

# Building on *Discovering Babel*: Plans for the Future

- New large sets of converted and new TEI XML P5 texts coming online
- Become a member of the *CLARIN SPF*
- Crosswalks in order to provide the original *TEI Headers* (with slight omissions) and *CMDI* metadata via OAI-PMH.
- Bringing the *British National Corpus* into the CLARIN realm
- As the web services/tools do not “come to the data” in the near future, we will want to provide at least some limited *search/query functionality on OTA content* as a web service – hence project with National Grid Service (NGS).
- This might also allow for *integration* into platforms for chaining web services, e.g. WebLicht

# The CLARIN Vision

---



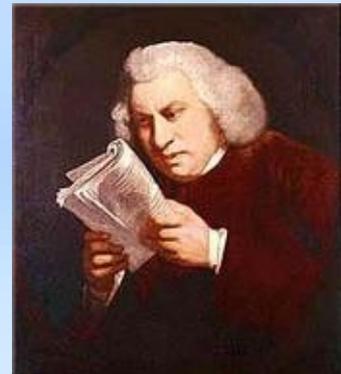
A researcher in Leeds from his desktop computer can:

- sign on with local authentication, and then:
- search for, find and obtain authorization to use corpora in Oxford, Prague and Bergen
- select the precise dataset to work on, and save that selection
- run semantic analysis tools from Budapest and statistical tools from Tübingen over the dataset
- use computational power from the local or national computing centre where necessary
- save the workflow and results of the analysis, and share those results with collaborators in Paris, Vienna and Zagreb
- discuss and iteratively adopt and re-run the analyses with collaborators

# Challenges for creators of resources

---

- **Standards**
- **Interoperability**
- **Sustainability**
- **Connecting and linking:** how can I facilitate research using my data along with other ones?
- **Creating a finished product:** when should I stop developing and make it available?
- **Impact:** how can I create the maximum impact with my data, in my academic field and beyond?
- How do I ensure that my resource is made available in relevant ways ***as a service***



# JISC

Thank you for your attention



**Oxford University Computing Services**

[www.oucs.ox.ac.uk](http://www.oucs.ox.ac.uk)